Problem Statement:

The restaurant landscape in New York City is very crowded, confusing, competitive for both business owners and customers. When deciding on a restaurant out of the plethora of options available no matter which neighborhood you are in, customer reviews play a deeply significant role in the final decision. According to a study from 2020, 90% of diners report researching reviews before picking a spot. These reviews build the reputation and brand image of anything being advertised online. Forbes magazine references a study revealing that 98% of customers "see reviews as an essential part of the decision-making process". In my personal experience, my friends and I always make sure that the customer rating of a restaurant is more than 4 stars on Google Reviews since we have endless options giving us no reason to pick anything below this standard, so I am excited to see how some of the factors I study ultimately impact the success of the restaurant.

This project relates to the lectures/papers we discussed since it will make the use of databases and data collection and web scraping. I will also be cleaning data and using data visualization techniques to better understand the data from these multiple factors in a visually appealing way. I will likely be able to build a prediction model and testing the data to allow the prediction of future success of restaurants based on the different factors I will be analyzing.

The objective of this project will be to analyze which factors drive positive and negative sentiment in customer reviews for restaurants in New York City. These insights will be important to customers for understanding key items to focus on and saving them time from looking at the factors that are not so influential. This analysis will be important in helping business owners in understanding which components of their restaurant are hurting and helping their rating. The factors I plan to mainly look at are neighborhood, pricing, cuisine, area density, capacity, and customer demographic which will help me determine which factors are the most important.

Data and Methods:

Sources:

- Yelp Fusion API: reviews, ratings, restaurant information
- Google Maps API: geospatial data, competitors in local area

Techniques and Methods:

- Sentiment analysis done using Natural Language Processing
    - VADER for sentiment scores
    - Assign positive or negative based on thresholds
- Top-n grouping
    - Sort and filter data based on the leading factors correlated with the most positive reviews
- Geospatial analysis
    - Cluster the restaurants using K-nearest to identify neighborhood trends

Data Preprocessing:

- Clean the review data using tokenizers, removing stopwords, and lemmatizing

Visualization:

- Create interactive graphs that can be used to display trends in sentiment with changing factors (such as increasing price or changing neighborhood)

Summaries:

- Provide fundamental summaries on which pricing has the highest reviews, locations with the highest reviews, etc.

Existing Issues:

- Platforms are isolated from each other, which makes it difficult to obtain a unified view of customer feedback.
- There is a lack of advanced analytics on the text of unstructured reviews, which prevents actionable insights.

Novelty:

- This project uses sentiment analysis and geospatial data
- Combines sentiment analysis, topic modeling, and geospatial data for a multidimensional analysis.
- Predictive insights for future reviews based on patterns and themes identified.

Related Works:

- Studies using Yelp datasets focus primarily on sentiment polarity without identifying granular factors that drive reviews.
- Health compliance studies fail to integrate customer sentiment as a contextual factor.

Evaluation Methods:

- We can use NLP Models: use accuracy and F1-score to validate sentiment analysis.
- Geospatial Clustering: measure clustering quality using k-means clustering.

Testing and Success Metrics:

- Success Metric 1: high accuracy for sentiment classification.
- Success Metric 2: actionable topics identified in at least 90% of reviews
- Success Metric 3: high user satisfaction with interactive graph usability